AD-A050 304 SOUTHEASTERN MASSACHUSETTS UNIV NORTH DARTMOUTH DEPT --ETC F/G 12/1
A NEW LOOK AT THE STATISTICAL PATTERN RECOGNITION.(U)
AUG 77 C H CHEN AFOSR-76-2951
UNCLASSIFIED EE-76-4 AFOSR-TR-78-0145 NL

1 OF 1
AD A050 304

END
DATE FILMED
3-78
DDC

AFOSR-TR- 78-0145

TECHNICAL REPORT SERIES IN INFORMATION SCIENCES
(Dr. C. H. Chen, Principal Investigator)

SOUTHEASTERN MASSACHUSETTS UNIVERSITY

ELECTRICAL ENGINEERING DEPARTMENT

NORTH DARTMOUTH, MASS. 02747 U.S.A.

Grant AFOSR 76-2951
IR No. EE-76-4

(2)

A NEW LOOK AT THE

STATISTICAL PATTERN RECOGNITION

by

C. H. Chen
Department of Electrical Engineering
Southeastern Massachusetts University
North Dartmouth, Massachusetts 02747

August 10, 1977

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(18) REPORT DOCUMENTATION PAGE (19)

READ INSTRUCTIONS BEFORE COMPLETING

1. REPORT NUMBER: AFOSR-TR-78-0145
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

(6) 4. TITLE (and Subtitle): A NEW LOOK AT THE STATISTICAL PATTERN RECOGNITION

(9) 5. TYPE OF REPORT & PERIOD COVERED: Interim rept.

6. PERFORMING ORG. REPORT NUMBER: (14) EE 76-4

(10) 7. AUTHOR(s): C. H. Chen

(15) 8. CONTRACT OR GRANT NUMBER(s): AFOSR-76-2951

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Southeastern Massachusetts University
Department of Electrical Engineering
North Dartmouth, MA 02747

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
(16) 61102F 2304/A2 (17)

11. CONTROLLING OFFICE NAME AND ADDRESS
(11) Air Force Office of Scientific Research/NM
Bolling AFB, DC 20332

12. REPORT DATE
(12) Aug 77

13. NUMBER OF PAGES
(12) 26 p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

It has now been 20 years since the publication of the original paper, "An Optimum Character Recognition System Using Decision Function" by C. K. Chow in IRE Trans. on Electronic Computer in 1957, in which he formulated pattern recognition as a problem of statistical decision theory. During the last two decades, statistical pattern recognition was well developed in theory and applications with the peak activity in the late sixties. The areas has now reached a farily saturated condition as its capability and limitations are well explored. The limitations are obvious:

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE → next page UNCLASSIFIED

407932

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract

the patterns are not characterized by the statistical information alone and many useful statistical properties cannot be fully developed with available mathematical statistics. The paper outlines important but unsolved problem areas in statistical pattern recognition and then takes a new and close look at some problems which are related to the finite sample size constraint. In an effort to bridge the gap between theory and practive, constructive solutions are provided for the problems: finite sample distance and information measures, finite sample nearest neighbor decision rule, contextual analysis, decision rules based on discrete and continuous measurements, and finite sample stochastic syntax analysis. It is concluded that there are still many challenging problems to be solved in statistical pattern recognition and every effort should be made such that the theory works well in practice.

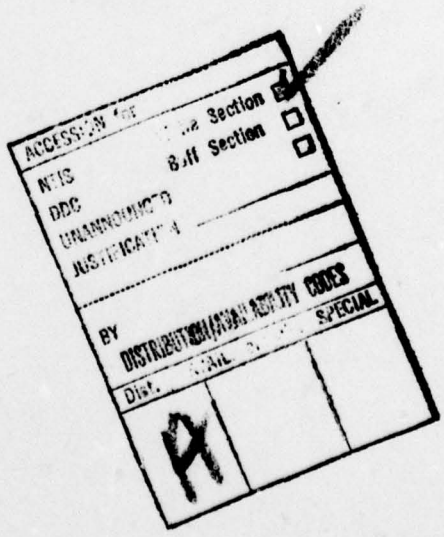
A NEW LOOK AT THE STATISTICAL PATTER RECOGNITION

C. H. CHEN

DEPARTMENT OF ELECTRICAL ENGINEERING
SOUTHEASTERN MASSACHUSETTS UNIVERSITY
NORTH DARTMOUTH, MASSACHUSETTS 02747

ABSTRACT

It has now been twenty years since the publication of the original paper, "An Optimum Character Recognition System Using Decision Function" by C. K. Chow in IRE Trans. on Electronic Computer in 1957, in which he formulated pattern recognition as a problem of statistical decision theory. During the last two decades, statistical pattern recognition was well developed in theory and applications with the peak activity in the late sixties. The area has now reached a fairly saturated condition as its capability and limitations are well explored. The limitations are obvious: the patterns are not characterized by the statistical information alone and many useful statistical properties cannot be fully developed with available mathematical statistics. The paper outlines important but unsolved problem areas in statistical pattern recognition and then takes a new and close look at some problems which are related to the finite sample size constraint. In an effort to bridge the gap between theory and practice, constructive solutions are provided for the problems: finite sample distance and information measures, finite sample nearest neighbor decision rule, contextual analysis, decision rules based on discrete and continuous measurements, and the finite sample stochastic syntax analysis. It is concluded that there are still many challenging problems to be solved in statistical pattern recognition and every effort should be made such that the theory works well in practice.



A New Look at the Statistical Pattern Recognition

1. Introduction

It has now been twenty years since the publication of the original paper, "An Optimum Character Recognition System Using Decision Function" by C.K. Chow in IRE Trans. on Electronic Computer in 1957, in which he first formulated pattern recognition as a problem of statistical decision theory. During the last two decades statistical pattern recognition was well developed in theory and applications with peak activity in late sixties. The area has now reached a fairly saturated condition as its capability and limitations are well explored. The limitations are obvious: the patterns are not characterized by statistical information alone and even some important statistical properties cannot be developed with available mathematical statistics. This situation has been quite typical with the application of every branch of mathematics. However for researchers new or old to this area there are still many challenging problems remaining to be solved. In [1], ten problem areas where the solutions are .. wanted are listed, not necessarily in the order of importance, as: feature extraction, nonstationary patterns, adaptive systems, learning complexity, finite sample size effects, computational recognition complexity, contextual analysis, optimum pattern recognizer, statistical and syntactic mixed model, and the automatic generation of recognition rules for complex patterns. It is hoped that good solutions to some of these problems will become available in the next decade.

It is not intended to survey the area, which is now quite broad, in this paper. Many books and articles have done this survey well. Instead the paper takes a new and close look at certain problems in statistical pattern recognition and offers some constructive solutions. Particular attention is given to bridging the wide gap between theory and practice, notably the problems of finite sample constraint.

II. Finite Sample Distance Measures

Distance measures are useful for feature selection and extraction and for error bounds of Bayes error probability (see e.g. [2], Chapter 4). They have been

extensively examined in recent years under the assumption of large sample size. In practice the sample size may be limited or small and many conclusions drawn under infinite sample assumption may not be valid under finite sample constraint [3]. The discussion here will be limited to Gaussian measurements for divergence and Bhattacharyya distance but can be easily extended to other cases.

Consider first the case of two univariate Gaussian densities with means m_1 and m_2 and the same variance σ^2. Last "^" denote the quantity evaluated by using the sample estimates. Then the difference in divergence between infinite sample and finite sample sizes is

$$\hat{J} - J = \frac{1}{\sigma^2}\left[(\hat{m}_1 - \hat{m}_2)^2 - (m_1 - m_2)^2\right] \tag{1}$$

where we have assumed that σ^2 is known. The expected value of the difference

$$E(\hat{J} - J) = \frac{1}{N_1} + \frac{1}{N_2} > 0 \tag{2}$$

is always positive where N_1 and N_2 denote the numbers of samples for classes 1 and 2 respectively. It is also assumed that all samples are statistically independent. The positive bias given by Eq. (2) indicates that the divergence evaluated by using a finite number of samples can lead to an over optiministic estimate of the error probability.

Next consider the univariate Gaussian densities with zero means and variances σ_1^2 and σ_2^2. The divergence based on the sample estimated parameters is

$$\hat{J} = \frac{\hat{\sigma}_1^2}{2\hat{\sigma}_2^2} + \frac{\hat{\sigma}_2^2}{2\hat{\sigma}_1^2} - 1 \tag{3}$$

The ratio $\omega = \hat{\sigma}_1^2/\hat{\sigma}_2^2$ has the F-distribution with (N_1, N_2) degrees of freedom. The expected error due to the finite sample size is

$$E(\hat{J} - J) = \frac{\sigma_1^2}{\sigma_2^2}\,\frac{1}{N_2 - 2} + \frac{\sigma_2^2}{\sigma_1^2}\,\frac{1}{N_1 - 2} \geq 0 \tag{4}$$

where the positive bias can be significant for small sample sizes.

The Bhattacharyya distance based on the sample estimated parameters is

$$\hat{B} = \frac{1}{2}\log\frac{1}{2}\left(\frac{\hat{\sigma}_1}{\hat{\sigma}_2} + \frac{\hat{\sigma}_2}{\hat{\sigma}_1}\right) = \frac{1}{4}\log\frac{(1+\omega)^2}{4\omega} \tag{5}$$

By using the Taylor series expansion of $\hat{B}$ with respect to the true value B, and retaining terms up to the second order in the expression, we obtain

$$E(\hat{B} - B) = \frac{\sigma_2^2 - \sigma_1^2}{4(\sigma_2^2 + \sigma_1^2)} \left(1 - \frac{2}{N_2 - 2}\right) + \frac{\sigma_2^4 + 2\sigma_1^2\sigma_2^2 - \sigma_1^4}{8(\sigma_2^2 + \sigma_1^2)^2} \left[1 - \frac{4}{N_2 - 2} + \frac{N_2^2(N_1 + 2)}{N_1(N_2 - 2)(N_2 - 4)}\right] \quad (6)$$

which is negative for $\frac{\sigma_1^2}{\sigma_2^2} \geq 1 + \sqrt{2}$ and positive otherwise. As the sample sizes approach ∞, the bias is not zero because of the series truncation. However, the sample size effect is evident from Eq. (6).

Now consider the multivariate Gaussian densities for p-dimensional measurements. Let $\bar{X}_1$ and $\bar{X}_2$ be the sample mean vectors corresponding to the true mean vectors μ_1 and μ_2 of classes 1 and 2 respectively. Also let S be the sample estimate of the common covariance matrix Σ given by

$$S = \frac{1}{N_1 + N_2 - 2} \left\{ \sum_{i=1}^{N_1} (X_i - \bar{X}_1)(X_i - \bar{X}_1)' + \sum_{i=N_1+1}^{N_1+N_2} (X_i - \bar{X}_2)(X_i - \bar{X}_2)' \right\} \quad (7)$$

where X is the vector measurement from either class 1 or class 2.

For infinite sample case the divergence is

$$J = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \quad (8)$$

which is the same as the Mahalanobis distance. The divergence using sample estimated parameters is

$$\hat{J} = (\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2) \quad (9)$$

where $E[S] = \Sigma$. The covariance matrix of $\bar{X}_1 - \bar{X}_2$ is

$$E[(\bar{X}_1 - \mu_1 - \bar{X}_2 + \mu_2)(\bar{X}_1 - \mu_1 - \bar{X}_2 + \mu_2)'] = \Sigma \left(\frac{1}{N_1} + \frac{1}{N_2}\right)$$

Let $k = \frac{1}{N_1} + \frac{1}{N_2}$. The random variable $\hat{J}/k$ has Hotelling's T^2 non-null distribution (see e.g. [4]) with $N_1 + N_2$ samples and p degree of freedom given by

$$H_p'\left(\frac{\hat{J}}{k} \middle| f, \frac{J}{k}\right) d\frac{\hat{J}}{k} = e^{-J/2k} \sum_{r=0}^{\infty} \frac{(J/2k)^r}{r!} \frac{1}{B\left(\frac{p}{2} + r, \frac{f - p + 1}{2}\right)} \frac{(\hat{J}/kf)^{(p/2)+r-1}}{(1 + \frac{\hat{J}}{kf})^{(1/2)(f+1)+r}} d\left(\frac{\hat{J}}{kf}\right), \quad \hat{J} > 0. \quad (10)$$

where f is the degree of freedom of S. By using the formula

$$\int_0^\infty \frac{x^{\mu-1}}{(1+x)^{\nu}}\,dx = B(\mu,\ \nu-\mu)$$

the expectation of J can be written as

$$E[\hat{J}] = \frac{kfp}{f-p+1} + \frac{fJ}{f-p+1} \tag{11}$$

For p = 1, then f = 1 aNd Eq. (11) is the same as Eq. (2). For p > 1, $E(\hat{J})$ differs from J not only by an additive term depending on sample size but also by a multiplicative constant independent of the sample sizes. The undesired effect of finite sample size is quite evident from Eq. (11). For equal mean but unequal covariance case the Bhattacharyya distance experiences the same effect as divergence since B = J/8.

For two multivariate Gaussian densities with zero mean vectors and covariance matrices Σ_1 and Σ_2 whose unbiased estimates are V_1 and V_2 respectively, the divergence based on sample estimated parameters is

$$\hat{J} = \frac{1}{2}\,\mathrm{tr}(V_1V_2^{-1} + V_2V_1^{-1}) - p \tag{12}$$

Since the measurements from the two classes are independent,

$$E[\hat{J}] = \frac{1}{2}\,\mathrm{tr}\{E(V_1)E(V_2^{-1}) + E(V_2)E(V_1^{-1})\} - p \tag{13}$$

Both V_i and V_i^{-1} follow the Wishart distributions with expectations

$$E(V_i) = \Sigma_i, \quad E(V_i^{-1}) = \frac{N_i}{N_i-p-1}\,\Sigma_i^{-1}, \quad i = 1,2 \tag{14}$$

Thus

$$E[\hat{J}] = \frac{1}{2}\,\mathrm{tr}\left(\frac{N_2}{N_2-p-1}\,\Sigma_1\Sigma_2^{-1} + \frac{N_1}{N_1-p-1}\,\Sigma_2\Sigma_1^{-1}\right) - p$$

$$= J + \frac{1}{2}\,\mathrm{tr}\left(\frac{p+1}{N_2-p-1}\,\Sigma_1\Sigma_2^{-1} + \frac{p+1}{N_1-p-1}\,\Sigma_2\Sigma_1^{-1}\right) \tag{15}$$

where the bias term coincides with Eq. (4) for p = 1, i.e. the univariate case.

The above discussion clearly illustrates the effect of finite sample on the bias of estimated distance measures. The variance of such estimates may also be determined. In general the estimated divergence has positive bias while the behavior of Bhattacharyya distance is less predictable.

III. Finite Sample Information Measures

For feature selection, more informative features result in low classification errors. However, if the sample size is limited, information measures estimated from samples may not be as effective. Consider the equivocation for m classes defined as

$$H = - E[\sum_{i=1}^{m} P(\omega_i/X) \log P(\omega_i/X)] \tag{16}$$

where $P(\omega_i/X) = P_i$ is the a posteriori probability of the ith class and the expectation is taken with respect to the space of X.

The sample-based equivocation using the estimated a posteriori probability P_i is

$$\hat{H} = - E[\sum_{i=1}^{m} \hat{P}_i \log \hat{P}_i] \tag{17}$$

Let θ_i be the parameter of the ith class, and $\hat{\theta}_i$ its estimate. Assume that the effect due to sample size is small so that we need consider only the first two terms in the Taylor series expansion of $\hat{P}_i$,

$$\hat{P}_i \simeq P_i + P_i'(\hat{\theta}_i - \theta_i) \tag{18}$$

where P_i' is the partial derivative of $\hat{P}_i$ with respect to $\hat{\theta}_i$ evaluated at $\hat{\theta}_i = \theta_i$. The difference between the estimated and true equivocations can be written as

$$\hat{H} - H \simeq E \sum_{i=1}^{m} [P_i'(\hat{\theta}_i - \theta_i)(1 + \log P_i) + P_i'(\hat{\theta}_i - \theta_i)^2] \tag{19}$$

which is still a function of $\hat{\theta}_i$. If $\hat{\theta}_i$ is an unbiased estimate of θ_i, then the expectation of the difference with respect to the estimated parameter depends only on the variance of $\hat{\theta}_i$ which is usually inversely proportional to the sample size. The variance of $\hat{H}$ given by $E(\hat{H} - E(\hat{H}))^2$ where the expectations are with respect to $\hat{\theta}_i$ can be shown to be proportional also to the variance of $\hat{\theta}_i$ or inversely proportional to the sample size. For the Renyi's information, it has been shown [5] asymptotically that the sample estimate of the information can have a variance of the order of the reciprocal of the squared sample size under certain condition. Thus the equivocation estimated under finite sample can be quite inaccurate.

IV. Finite Sample Nearest-Neighbor Decision Rule

The nearest-neighbor decision rule (NNDR) is attractive in the sense that the NN-risk is upper bounded by twice the Bayes risk when the sample size approaches infinity. For a given sample size the 1-NNDR is uniformly better than the k_n-NNDR. The small sample NNDR performance has been considered for the restrictive cases: Fix and Hodges [6] investigated the small sample performance for k_n-NNDR for univariate and bivariate normal distributions; Levine et.al. [7] showed that the performance for small sample sets from uniform distributions is close to its asymptotic value. For multivariate Gaussian densities and allowing the sample size to increase with k, it is shown [8] numerically that the k-NNDR has a very close performance as the Bayes linear discriminant analysis. This result is significant in the sense that under finite sample condition the NNDR is comparable to the Bayes rule using estimated parameters. For a given set of n samples with known classification, it would be more meaningful to compare different decision rules using the n samples rather than to compare with Bayes rule under infinite sample size assumption. For the Gaussian assumption the NNDR is very competitive with other decision rules based on the results of [8].

In practical use of nearest-neighbor rule, the large number of samples would require large amount of storage and computation. Methods for reducing the computation requirement and editing the samples have been considered. It has been established experimentally that there is always a small subset of good learning samples that dominate the performance. In other words the performance would be insensitive to sample size for good quality neighbors. This idea is somewhat similar to the edited NNDR which attempts to eliminate samples on the wrong side of decision boundary.

The fundamental question whether the Euclidean distance is most effective in NNDR has not been resolved. Experimental results based on some weighted Euclidean distance have indicated better recognition performance than on Euclidean distance. If the samples are close to be normally distributed, a better

distance computation makes use of the covariance matrix for each class, i.e. for each neighbor $P^{(i)}$ belonging to the ith class, compute the squared distance

$$(x - P^{(i)})' V_i^{-1} (x - P^{(i)}) = d_i^2 \tag{20}$$

and choose the class which provides the minimum d_i^2. The Euclidean distance may be considered as a special case of Eq. (20) by setting $V_i = I$. If the nearest-neighbor is considered as the reference point of a class then for the minimum distance classifier provided by the Euclidean distance NNDR, there exists a linear decision function, obtained by using common convariance matrix, which is at least as good according to the classification theory (Chapter 2 of [2]). If the covariance matrices are quite unequal then the quadratic classifier provided by Eq. (20) is better than the linear classifier. Exactly how much better is a question better answered experimentally. Recent investigation with the seismic data [9] has shown that the modified distance given by Eq. (20) provides more than 15% improvement in the percentage correct recognition than the Euclidean distance NNDR. Of course the sample size must be large enough to calculate the covariance matrices accurately.

In summary, the NNDR is an effective and reliable decision rule for finite sample size condition. Especially for small sample size when a good estimate of parametric density is not available, the NNDR should be used.

V. Contextual Analysis

A major weakness of statistical pattern recognition is the difficulty to take the contextual relations into account in the recognition process. Character recognition is not considered here as it requires somewhat different contextual analysis [10]. An imagery pattern is rich in contextual information part of which is statistical in nature. A formal statistical approach to this problem is the compound decision theory. The finite sample constraint in digital imagery patterns is caused by limited number of images and the limitation in spatial resolution. In image interpretation and classification, an image is usually

partitioned into a number of subimages. A vector measurement may be taken from each subimage. By assuming dependence on the nearest four neighbor subimages, the compound decision rule is to choo-se the class which maximizes (ω_k, ω_j = 1,2,... m)

$$P(X_k/\omega_k)P(\omega_k) \prod_{j=1}^{4} \sum_{\omega_j} P(X_j/\omega_j)P(\omega_j/\omega_k) \tag{21}$$

which is adapted from the last equation on page 201 of [2]. Here X_k is the vector measurement for the kth subimage. Notice that the part of the expression outside of the product sign is identical to that used in a simple maximum likelihood decision rule without considering neighbor subimages at all. The product t-erm represents the contextual information for the kth subimage. Each multiplier in the product term represents the contextual contribution from an adjacent neighbor subimage. By rewriting the multiplier as

$$\sum_{\omega_j} P(X_j/\omega_j)P(\omega_j) \frac{P(\omega_j/\omega_k)}{P(\omega_j)} \tag{22}$$

it is seen that computationally this is a weighted histogram of the subimage j, with each class ω_j being weighted by the factor $P(\omega_j/\omega_k)/P(\omega_j)$ which reflects the dependence between two states of nature for two adjacent subimages k and j. The accuracy of the weighted histogram is related to the performance of compound decision rule given by Eq. (21). The sampling distribution of the weighted histogram for q quantization levels and a total of n pixels for the jth subimage is

$$\frac{\Gamma(n+q)}{\Gamma(r_{ij}+1)\cdots\Gamma(r_{qj}+1)} \prod_{i=1}^{q} P_{ij}^{r_{ij}} \tag{23}$$

where r_{ij} is the number of pixels belonging to the ith quantization level. The Bayes estimate of P_{ij}, the fractional number of pixels for the ith level is

$$\hat{P}_{ij} = \frac{r_{ij}+1}{n+q} \tag{24}$$

By using Eqs. (21)-(24) and following the analysis of [11], an average probability of correct recognition for the subimage k can be determined as a function of sample size (i.e. the number of pixels n), and the number of quantization levels q.

If we consider two classes only such as object ($\omega_k = 1$) and background ($\omega_k = 2$), then the effect of contextual dependence appears as a multiplicative factor in the likelihood ratio. A suboptimal but much simpler scheme to determine such factor is to compute

$$\prod_{j=1}^{4} \frac{P(\omega_j = 1/\omega_k = 1)/P(\omega_j = 1) + P(\omega_j = 2/\omega_k = 1)/P(\omega_j = 2)}{P(\omega_j = 1/\omega_k = 2)/P(\omega_j = 1) + P(\omega_j = 2/\omega_k = 2)/P(\omega_j = 2)} \tag{25}$$

as the histograms computed for all four neighbor subimages tend to cancel out in the numerator and denominator. Other simple ways to profitably utilize the statistical contextual information in image analysis should also be examined both theoretically and experimentally.

VI. Decision Rules Based on Discrete and Continuous Measurements

Most pattern recognition work assumes either discrete or continuous measurements (including measurement quantized from the continuous one). In image recognition, it is possible to tentatively assign each subimage to one of several possible classes, which is a discrete quantity, while the actual measurement of the subimage is continuous. In the decision tree framework, an overall classification of the image may be made by using all the informations on each subimage including the preliminary decision made on it. Similar situation arises in medical diagnosis in which the final diagnosis depends on decisions made on some tests and other continuous measurements.

Recently, Krzanowski [12] considered the use of Fisher's linear discriminant function for classification with a set of p continuous and q binary variables. His work is immediately applicable to medical data. From the information provided by the discrete variable, a likelihood ratio is formed on the continuous variable and compared with a threshold determined by the discrete variable. For image analysis, a decision or interpretation has to be made on an image containing a number of subimages on which individual decision may be made first. A bottom-up decision tree may be established to reach the best final decision. Inconsistent decisions between two neighbor subimages may indicate the existence of an object

boundary or an incorrect decision on one of them. Backtracking or error correction mechanism may be added to the decision making process. The size of the subimage should be chosen so that it is much smaller than the object size. The decision process can be summarized as follows:

Step 1. Starting from the first subimage, compare its decision with all eight neighbors. Proceed next with one of the subimage of same decision. If no consistent decision is available, proceed with any neighbor.

Step 2. Examine the second subimage in the same manner as Step 1. Repeat the step as many times as needed until returning to the first subimage with closed boundary. The desired object is located.

Step 3. If a closed boundary is not available after search and merge in Steps 1 and 2, then decision is made that the image does not contain the object.

Obviously other decision tree procedures can be established (see e.g. [13]) for the same objective. These procedures are much easier to implement than the use of "one shot" compound decision function.

VII. Finite Sample Stochastic Syntax Analysis

The production probabilities in stochastic syntax analysis [14] are usually estimated from a set of distinct sample strings by frequency ratio. The limited string sample size is a source of error in estimation and the final classification performance. The error accumulates as a sequence of production rules is applied. The nonmonotonic relation between confidence for $\hat{p}_{ij}$ and sample size (page 181 of [14]) is rather unexpected. To simplify the analysis, assume that M distinct production rules have to be applied to complete a parse. Let e_{ij} be the difference between the estimated and true production probabilities, i.e.

$$e_{ij} = \hat{p}_{ij} - p_{ij} \tag{26}$$

Then $E(e_{ij}) = 0$, cov. $(e_{ij}e_{ik}) = - p_{ij}p_{ik}/\bar{n}$, $j \neq k$ where $\bar{n} = \sum_j n_{ij}$ with n_{ij} defined by Eq. (6.4) of [14]. $\hat{p}_{ij}$ approaches p_{ij} as the number of sample strings t, i.e. the sample size, approaches infinity. The value $\bar{n}$ is proportional to t.

The likelihood function for the grammar of a given class is

$$\prod_{j=1}^{M} \hat{p}_{ij} = \prod_{j=1}^{M} (p_{ij} + e_{ij})$$

which has the expected value

$$E\left(\prod_{j=1}^{M} \hat{p}_{ij}\right) = \prod_{j=1}^{M} p_{ij} - \frac{M(M-1)}{\bar{n}} \prod_{j=1}^{M} p_{ij} + \text{higher order terms} \qquad (27)$$

If we ignore the higher order terms then the likelihood function based on estimated production probabilities is expected to be off from the true value by an amount inversely proportional to the sample size and proportional to M^2. For long string the accuracy of the likelihood function may thus be very poor. The variance of the likelihood function can also be determined. It appears that the only way to reduce the finite sample effect is to increase the sample size.

VIII. Concluding Remarks

This paper has examined some current problems in statistical pattern recognition especially the effects of finite sample size, which cause the gap between theory and practice in pattern recognition. When the effects are monotonic then the best way to reduce such effect is probably by increasing the sample size. There are many other problems, as listed in Section I, in statistical pattern recognition which remain to be studied also. Thus we believe the area should remain to be an active one for researchers.

References

1. C.H. Chen, "Statistical pattern recognition - review and outlook", IEEE Systems, Man and Cybernetics Review Special Issue on Current Perspectives of Pattern Recognition, August 1977.

2. C.H. Chen, "Statistical Pattern Recognition", Hayden Book Co., February 1973.

3. C.H. Chen, "On statistical and structural feature extraction", in "Pattern Recognition and Artificial Intelligence" edited by C.H. Chen, Academic Press, Inc., 1976.

4. A.M. Kshirsagar, "Multivariate Analysis", Marcel Dekker, Inc., New York, 1972.

5. J. Zvarova, "On asymptotic behavior of a sample estimator of Renyi's informations of order α", Trans. of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, 1973.

6. E.Fix and J.L Hodges, Jr., "Discriminatory analysis: small sample performance", USAF School of Aviation Medicine, Randolph Field, Texas, Project 21-49-004, Rept. 11, August 1952.

7. A. Levine, L. Lustick and B. Saltzberg, "The nearest-neighbor rule for small samples drawn from uniform distributions", IEEE Trans. on Information Theory, Vol. IT-19, No. 5, pp. 697-699, September 1973.

8. C.H. Chen, "Performance bounds of a class of sample-based classification procedures", Proc. of Pattern Recognition and Image Processing Conference, June 1977.

9. C.H. Chen, "On digital signal modeling and classification with teleseismic data", TR EE-75-5, August 1977.

10. G.T, Toussaint, "The use of context in pattern recognition", Proc. of Pattern Recognition and Image Processing Conference, June 1977.

11. B. Chandrasekaran and A.K. Jain, "Quantization complexity and independent measurements", IEEE Trans. on Computers, pp. 102-106, January 1974.

12. W.J. Krzanowski, "The performance of Fisher's linear discriminant function under non-optimal conditions", Technometrics, Vol. 19, No. 2, pp. 191-200, May 1977.

13. C.L. Wu, "The decision tree approach to classification", Ph.D. thesis, Purdue University, May 1975.

14. K.S. Fu, "Syntactic Methods in Pattern Recognition", Chapter 6, Academic Press, Inc., 1974.